Mini-Chapter — Modeling Time Series & Classification

1. Problem Types at a Glance

- **Regression:** predict a continuous number (e.g., next-day return).
- **Classification:** predict a class/label (e.g., up/down, fraud/not).
- **Time Series:** data where order matters; past influences future; splits must be time-aware.

2. Why Returns, Not Prices

Raw prices trend and scale over time, risking spurious relationships. **Returns** (or **log-returns**) are closer to stationary and comparable across levels. In finance, using returns is also natural for risk/utility framing.

3. Features for Time Series

- **Lags:** `ret(t-1)`, `ret(t-5)` catch momentum/reversal.
- **Rolling means:** smooth noise to reveal trend.
- **Rolling volatility (std):** proxy for **risk** state.
- **Momentum:** cumulative return over a window.
- **Don't leak:** only use information available at time t to predict $t+1$.

4. Splitting the Data

- **Random split** breaks the temporal order → optimistic metrics.
- **Time-aware split** (train: past; test: future) or `TimeSeriesSplit` mimics reality.
- Use the last 20–30% as test for simple baselines.

5. Classification Intuition

- **Logistic regression** outputs probabilities; threshold converts to class.
- **Decision trees** split feature space into regions; intuitive but can overfit.
- **Imbalance:** Accuracy can hide failures—use **precision/recall/F1**.

6. Pipelines for Reproducibility

- Bundle steps: scaling, feature gen, model.
- Fit transforms on train only; reuse on test.
- A stable pipeline is the backbone of your modeling repo.

7. Evaluating Usefully

- Forecasting: **MAE/RMSE**, plot predicted vs true, inspect **residuals**.
- Classification: **precision/recall/F1**, **confusion matrix**; align metrics to **costs** (false positives/negatives).

8. Failure Modes to Watch

- Regime shifts: model trained in one volatility regime fails in another.
- Leakage via `.shift(-1)` or merging with future signals.

- Over-smoothing with large windows; delayed reaction.
- Misusing accuracy with imbalanced outcomes.

9. Where This Leads

In later modules: ARIMA/VAR/GARCH for time series; tuned tree ensembles and calibrated classifiers; cross-validation that respects time; model risk and backtesting discipline.